(1)

Final Technical Report
on Phase I SBIR Study
on "Semi-Automated Speech Transcription System"
at Dragon Systems

Semi-Automated Speech Transcription System Study

Steven Wegmann
August 1994

Sponsored by
Advanced Research Projects Agency (DOD)

Defense Small Business Innovative Research Program

ARPA Order No. 5916

Issue by U.S. Army Missile Command Under

Contract # DAAH01-93-C-R101

Principal Investigator: Janet Baker

Name of Contractor: Dragon Systems Inc.
Business Address: 320 Nevada Street, Newton, MA 02160.
Telephone: (617) 965 5200
Electronic Mail: melvyn@dragonsys.com

Effective Date of Contract: 1-29-93
Contract Expiration Date: 8-31-93
Reporting Period: 1-29-93 to 8-31-94

Technical Contacts: Janet Baker and Larry Gillick
Administrative Contact: Harry Taylor

Disclaimer

"The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Advanced Research Projects Agency or the U.S. Government."

Abstract

This report describes preliminary explorations towards the design of a semi-automatic transcription system. Current transcription practices were studied and are described in this report. The promising results of several speech recognition experiments as well as a topic identification experiment, all performed on broadcast data, are reported. These experiments were designed to gauge the quality of speech recognition on broadcast data and to explore possible uses of a continuous speech recognizer in a semi-automatic transcription system. Possible future directions for research are also reported.

Dragon Systems Proprietary Claims

The experiments reported here have been based in part on proprietary technology which Dragon Systems has already developed at private expense. This proprietary technology includes techniques for greatly reducing the amount of computation required to recognize large vocabularies and is covered by US patents 4783803, 4803729, 4805218, 4805219, 4829578, 4829576, 4837831, 4866778, 4903305, and 5027406, and other pending patent applications. Since Dragon Systems is a small business, new technology developed under this contract is provided with Government Use Rights with regard to new patents and copyrights. Per contractual agreement, Dragon Systems software developed at private expense or based on existing inventions developed at private expense will be provided with restricted rights.

1. Introduction

The main task of this phase of the project was to conduct preliminary explorations towards the design of a semi-automatic transcription system. Although fully automatic recognition and transcription of unrestricted continuous speech is beyond the current state of the art, possible uses of speech recognition technology in the transcription process will be explored in this report.

The first task of the project was to understand how the current transcription process works. Direct communication and interaction with skilled transcriptionists and captioning centers took place, the results of which are summarized in sections two and three.

Next, broadcast data were obtained from the British Broadcasting Corporation. These data were processed so that they could be used in several recognition experiments, of a very preliminary nature, that tested the quality of automatic speech recognition and explored some possible uses of a continuous speech recognizer in a semi-automatic transcription system. These experiments are described in section four.

Section five of this report consists of a discussion of possible future directions for this research.

2. Survey of Available Materials

Dragon Systems contacted the Subtitling Department of the British Broadcasting Corporation, the National Captioning Center, the National Captioning Institute, and The Caption Center at the WGBH Educational Foundation. Information and materials concerning the transcription process and its uses were obtained and studied from all four sources. For example, the BBC, the National Captioning Institute, and WGBH gave extensive tours of their facilities. The National Captioning Institute gave a presentation of current and potential educational uses of captioned broadcast data. In addition, the BBC and WGBH supplied information about their current captioning practices.

Obtaining transcribed broadcast data from these sources was a different matter. For various reasons, only the BBC was willing to supply transcribed broadcast data at this time. The BBC was especially interested in this project: they generously provided videotapes of television programs with the captions displayed (with

examples of the two styles of captioning described below), hard copy and binary samples of captions, as well as the verbatim transcripts of two news programs.

In addition to the material made available through these contacts, there is, of course, a huge quantity of data that is broadcast every day through the airwaves and cable, which could be recorded at cost and used for the experimental purposes described in this report. Transcripts of some of this data, particularly talk shows, are available for modest fees from the networks that produce them. Data that does not have transcripts available could be transcribed on a contract basis.

3. The Problem of Transcribing Broadcast Data

Today, broadcast data are processed in two main ways: an off-line, non-real-time mode and in a on-line, real-time mode. In each of these modes, there are two possible types of output: verbatim transcripts or edited transcripts (captions). Both modes of transcription are very labor intensive tasks requiring highly skilled typists. The familiar task of transcribing speech produced for dictation provides a useful benchmark for assessing the difficulty of transcribing broadcast data. In speech produced for the purpose of dictation, the speaking rate is typically between 60 - 90 words per minute. Skilled transcribers, who are familiar with the dictated material, normally require 4 - 6 times real-time to transcribe this sort of data. However, in conversational speech the speaking rate is typically between 150 - 230+ words per minute, with reputed bursts of 300 words per minute.

Not surprisingly, then, most broadcast data are transcribed after the fact in a non-real-time mode with the aim of producing either verbatim transcripts or captions. By using specialized equipment that allows them to vary the playback of the audio data with foot pedals, highly skilled typists (typing at about 100 words per minute) produce transcriptions at a rate of about 3.5 times real-time.

Some broadcast data, e.g. important live news broadcasts, are transcribed in close to real-time with lags of up to 5 seconds. Fast, highly skilled typists use court stenography machines to accomplish this task with reasonable accuracy even with speaking rates as high as 220+ words per minute. The error rates in the resulting transcriptions vary tremendously depending on the speaking rates, the uniformity of the material being transcribed, and the availability of

tools such as custom on-line dictionaries. These highly skilled on-line transcribers are in short supply and, as a consequence, are even more expensive to use than the people who do the off-line transcription.

As a matter of nomenclature, Europeans tend to refer to transcripts of spoken materials which are formatted for display on a screen as "subtitles", whereas Americans refer to these as "captions". With all due apologies to our British colleagues, we will use the American nomenclature in this report. The captions that are produced are of two main types (see [WGBH]): pop-on captions (similar to movie subtitles) that appear on the television screen one by one, and roll-up captions in which the words continuously scroll by in a window of text at the bottom of the screen. Pop-on captions are the most labor intensive to produce since they are edited from transcripts to appear at an appropriate reading speed, they are synchronized to the audio of the television program, and they have marks to indicate a change in speaker. In order to produce good pop-on captions, verbatim transcripts are usually edited down. For readability, it is necessary to reduce the amount of speech displayed when the speaking rate is fast. Some transcription practices (e.g. the National Captioning Institute) impose strict limits on the number of words displayed in a caption, depending on the age or reading ability of the intended audience. Other practices (e.g. the BBC) emphasize displaying transcripts as fully as practical in the interests of providing more complete and faithful renderings. Roll-up captions are easier to produce since the transcripts are just scrolled on the screen, usually with minimal editing. Most of the output of on-line transcription is displayed using roll-up captions, but roll-up captions are also used to display the output of off-line transcription when the issues of time or added expense preclude the preparation of pop-on captions or where displaying the most complete rendering is desired.

Overall, the current techniques used for transcribing broadcast data are labor intensive and require specialized equipment, with the added drawback that very few people have the necessary training to do this highly skilled labor.

The current state of the art in speech recognition technology does not come close to being able to automatically transcribe conversational speech in real time at anywhere near reasonable error rates. But a state-of-the-art speech recognizer could be part of a semi-automated transcription system which could dramatically reduce the

difficulty of the task of transcription by

1) reducing the necessary typing skills of the human operator, by having the operator and the speech recognizer work together at producing the transcript.

2) taking over or simplifying the control of the playback of the audio data, in the case of off-line transcription, again reducing the necessary skill level of the human operator.

3) automatically loading domain specific, on-line dictionaries based upon the decisions of a topic identification system.

The preparation of captions could also be simplified somewhat by using speech recognition technology to aid in their synchronization to the audio tract.

Given the current state of the art, most of the benefits of such a system would apply to the off-line transcription of speech, but, as the state of the art in speech recognition improves, gradual progress on the problem of on-line transcription could be made and the distinction between on-line and off-line transcription would, in effect, vanish.

Such a semi-automated transcription system could, by reducing the necessary skill level of the human operator, reduce the costs associated with producing transcriptions thereby making them more attractive to produce. This would benefit the hearing impaired community by making more television programming available to them. Captioned broadcasts are also being used as a learning aid for students with learning disabilities, as well as students for whom English (or another language) is a second language, so both of these groups would also benefit from any increase in data at lower costs.

4. The Experiments

There are seminal questions to be answered before any serious development can begin on a semi-automatic transcription system, among which are "what kind of broadcast data are readily available?", "what are the issues surrounding its efficient collection?", "what are these data like acoustically?", and "how well do they respond to standard speech recognition techniques?" The experiments described below were meant to get initial readings on these questions as well as to explore some simple techniques that might be useful in a semi-

automatic transcription system.

First the data were collected and processed, a labor intensive task with elements in common with the current procedure for transcribing broadcast data off-line. Next, two simple experiments were performed to get an initial sense of the usability of the collected data: a segmentation experiment and a recognition experiment, with some adaptation, were run using standard speech recognition tools and models in conjunction with a state-of-the-art research recognizer for large vocabulary continuous speech recognition. Then, an interesting experiment was run in order to test speaker adaptive recognition on this data. The last experiment was a topic identification experiment. Text from the output of a speech recognizer was run through a pre-existing, in-house topic identification research system to assess the reliability with which useful information about the topic of the speech could be extracted.

Data Collection and Processing

The BBC generously provided three television programs on videotape: a news program, a science program called Horizon, and a children's program called Landmarks. The BBC also provided a hard copy transcript of the news program and hard copy and binary versions of the actual subtitles that their close captioning system produced for the other two programs.

The recognition experiments that were run require speech data in a machine readable standard wavefile format, plus companion text files with the transcripts of what was spoken which are used for scoring during recognition and for alignment during training. The first step, then, was to transfer the audio portion of the videotapes to the computer using standard tools for converting the data into wavefiles. In the case of the news program, which did not come with a machine readable transcript, only the speech of the news announcers was recorded. These data were then, through the use of a wavefile editor, cut into smaller units corresponding to the sentences spoken by the announcers. The corresponding transcripts were created by hand. The other two programs, which did come with machine readable subtitles, were processed somewhat differently. The speech portion of the audio data was cut into smaller units corresponding as much as possible to the supplied subtitles using a wavefile editor, then the subtitles, after being extracted from the provided files, were edited to match what was actually spoken. The result of this processing, counting all three television programs, was 64 minutes of speech

from 28 speakers split up into 744 pairs of audio data and matching transcripts. The average duration of an utterance was 5 seconds. A total of 51 of the utterances came from the news program, 521 came from the science program, and 172 came from the children's television program.

By far the most time consuming part of the data collection process was the hand editing of the speech data into the sentence-like chunks of the subtitles. Furthermore, 67% of the subtitles needed modifying to match precisely what was spoken.

Dragon Systems' Speech Recognizer

The large vocabulary continuous speech recognizer that was used for all of the experiments described in this report was developed by Dragon Systems in conjunction the US Government's ARPA (Advanced Research Program Agency) SLS (Spoken Language System) program and was trained on the Wall Street Journal task (described in [recog1], [recog2] and [recog3]). This speech recognizer is a time-synchronous hidden Markov model based system. It requires a set of acoustic models, and a language model. Before recognition, the acoustic data need to be run through standard in-house signal processing which produces 32 acoustic parameters (8 spectral, 12 cepstral, and 12 cepstral difference parameters) in 10ms frames. An IMELDA transform ([IMELDA]), that was constructed via linear discriminant analysis, was applied to the acoustic parameters in order to produce a less highly correlated set of 16 acoustic parameters. Transcripts were used to score the recognized text and to adapt models when in adaptation mode.

It may be useful to compare the recognizer performance in the experiments described below with some performance benchmarks taken from previous experiments performed under the SLS program. This recognizer, with a 5K vocabulary and the gender independent acoustic models used in the experiments on the BBC data, obtained a 13.2% word error rate (86.8% word accuracy rate) on the November 1993 5K evaluation test data (this experiment was only used for in-house development purposes). A 5K vocabulary was not big enough for these experiments, since there were too many out of vocabulary words in the BBC data, so a 20K vocabulary was used for the experiments described below. With a 20K vocabulary and gender independent acoustic models similar to those used here (the acoustic models used in the experiments on the BBC data model half as many triphones), the recognizer ran with a 27.2% word error rate (72.8%

word accuracy rate) on the official WSJ1 20K development test. These models were chosen for use in the experiments described below because, being simpler, they could adapt more quickly to the limited training data available. They are not Dragon's best performing models. Dragon's best performance in the November 1993 WSJ 20K evaluation test was a 19.1% error rate (80.9% word accuracy rate) ([recog3]).

A priori, it was conjectured that the performance of this recognizer on broadcast data would lie somewhere between the results described above and the results that Dragon has obtained during recognition of telephone conversations taken from the SWITCHBOARD corpus (a standard speech corpus consisting of recorded telephone conversations on fixed topics collected at Texas Instruments and produced on CD-ROM by NIST (National Institute for Standards and Technology, formerly National Bureau of Standards)). This conjecture was made since the acoustic qualities of broadcast data lie somewhere between the extremes of the quality of Wall Street Journal data (very high) and SWITCHBOARD data (poor), but also because the type of speech in broadcast data is also somewhere between extremes of read text (the Wall Street Journal data) and spontaneous speech (SWITCHBOARD data). Dragon Systems' SWITCHBOARD recognizer obtained a 22% word correct rate on the December 1992 evaluation test (see [topic]). Recent, but very preliminary, development efforts have resulted in the SWITCHBOARD recognizer obtaining a 32.5% word accuracy rate.

These performance benchmarks, along with the best result from the experiments described below (experiment 2c), are presented in table 1.

Segmentation Experiment

Before proceeding with the main recognition experiments a small segmentation experiment was run to get a general sense of what the data were like. Given the phonetic spelling of all the words in the transcript of the speech, a segmentation is a labeling of each frame of the acoustic data by the phonetic element that was being spoken at that moment. A speech recognizer can be used to produce a segmentation automatically. Dragon's speech recognizer, when in segmentation mode, produces output that can be used by various in-house tools to create and view a spectogram of the speech with the phonetic boundaries clearly marked. The utterances from the news

Table 1.

When comparing recognizer performance, bear in mind that WSJ acoustic data are collected using a high quality, noise cancelling microphone, BBC acoustic data are from the audio track of a videotape, and SWB data are collected from telephone speech.

Recog[illegible]zer Task	Recognizer Accuracy (Word accuracy rate)	Amount of Training Data	Date
WSJ 5K	86.8%	60 hours	April 1994
WSJ 20K	72.8%	60 hours	April 1994
BBC 20K (experiment 2c)	46%	1 hour	July 1994

The acoustics models in this group are comparable. The acoustic models used when recognizing the BBC data were adapted from the same acoustic models used in the WSJ 5K test. The WSJ 20K acoustic models are very similar (but they have twice as many triphones than the WSJ 5k models have). The original WSJ models are speaker and gender independent, but in the BBC experiment these models were incrementally adapted, on 14 minutes of speech data, to a single male speaker.

WSJ 20K	80.9%	60 hours	November 1993

This result is Dragon Sytems' best official evaluation performance on the WSJ 20K task. The acoustic models are speaker independent, but gender dependent.

SWB 8.4K	22%	9 hours	December 1992
SWB 2.4K	32.5%	9 hours	June 1994

The first SWITCHBOARD result is the word correct rate, which is generally greater than word accuracy since

word correct rate = (#symbols -#subsitutions -#deletions)/#symbols

while

word accuracy rate = (#symbols -#subsitutions -#deletions -#insertions)/#symbols.

The first SWITCHBOARD recognizer used speaker independent, gender dependent acoustic models, while the second used speaker and gender independent acoustics models.

program were chosen to be segmented, because they formed a manageable sized set for viewing and they also seemed typical of good quality utterances, i.e. the sort that good models could be built from. The utterances were segmented using simple acoustic models built from TIMIT data (the TIMIT corpus is a standard, DARPA sponsored speech corpus, which was recorded at Texas Instruments, transcribed at MIT and prepared for CD-ROM at NIST, consisting of high quality acoustic data which has been hand-labeled by expert spectogram readers). The resulting segmentations were viewed by in-house experts who graded them "excellent" overall.

While this is a very subjective result, it is interesting for two reasons. The first is that a necessary ingredient for building acoustic models for broadcast data is good initial segmentations of a large quantity of data. The above result suggests that the initial segmentations could be produced automatically using the TIMIT models, rather than by hand. The second is that segmentations are not only useful for training new models: they can also be used to synchronize text to speech. For example, during the transcription process a speech recognizer in segmentation mode could keep the speech being played back synchronized with the point where the transcriber is typing.

Recognition Experiments

The 64 minutes of data collected were considered insufficient to build acoustic models from scratch, so instead acoustic models built for the DARPA SLS Wall Street Journal task were adapted to this new domain. Several caveats, which must be kept in mind when evaluating the results of these experiments, are in order regarding the suitability of these models for this task. The most important caveat concerns the noise levels of the BBC data. Wall Street Journal acoustic data are gathered from speakers reading Wall Street Journal articles into a high quality, noise canceling microphone, so the utterances have very little background noise and the channel characteristics of the data vary very little across the corpus. In contrast, the quality of the audio data taken from the videotape varies tremendously from utterance to utterance. Most of the utterances have some noise, most often music, in the background. Traffic noises, rumbling and speech also occur frequently in the background. Also, the BBC data have been obtained from a unknown number of different microphones of unknown quality, so the channel characteristics are very different from those of the Wall Street

Journal data, a difference which degrades recognition performance. Another difference to bear in mind is that the majority of speakers in the BBC data are British, unlike the Wall Street Journal speakers who are all American. Lastly, the majority of speakers in the BBC data are speaking spontaneously, compared to the speakers in the Wall Street Journal data who are reading text. To compensate in part for these acoustic differences the acoustic models were adapted in various ways, as will be described in the experiments below.

In addition to acoustic models, the recognizer requires a language model. Because there was insufficient data to build a language model for this domain, a language model built for the DARPA SLS Wall Street Journal task was used. This models written text, in particular Wall Street Journal text, not natural speech. While the BBC data do include some narrators who read from scripts, the scripts were certainly not taken from The Wall Street Journal! The Wall Street Journal language model was even less appropriate for the task of recognizing the majority of speakers who were speaking naturally in response to questions and on subjects ranging well beyond those covered in the Wall Street Journal domain.

Experiment 1

The purpose of this experiment was to get some idea of how well standard Wall Street Journal models work on the BBC data. The test data were partitioned into two sets: a test set consisting of 200 randomly chosen sentences was used for recognition and the training set consisting of the remaining 544 sentences was used to adapt the acoustic models to this domain. The experiment compared recognition of the test data before and after adapting the Wall Street Journal acoustic models. The word error rate before adaptation was 83%. The word error rate dropped to 71% after adaptation.

The results of this experiment say more about the limits of adapting acoustic models with such sparse data than anything else. The BBC data is a very small set of data with widely varying acoustic characteristics. When the Wall Street Journal acoustic models were adapted to this data there were many acoustic differences that needed adjusting but too few examples available to effectively make the adjustments. In other words, to get significantly better recognition performance, with a training set of this limited size, a set with more uniform acoustic characteristics should be used to adapt the acoustic models. The next experiment attempts to do that by focusing on a single speaker.

Experiment 2

One possible scenario for an adaptive, semi-automatic transcription system is that the user corrects the output of a large vocabulary recognizer, with the recognizer adapting its models based upon the user's corrections. A more automated process would simply adapt the models based on the (possibly errorful) recognized transcriptions. The purpose of this experiment was to explore how a speaker adaptive system might work.

The speaker with the most data in the BBC collection was the narrator from the Horizon program. His 170 utterances, approximately 14 minutes of data, were set aside as the test set, while the remaining data were used to adapt the Wall Street Journal acoustic models to the BBC domain. Three recognition tests were then run:

a) Recognition of the test sentences without doing speaker adaptation. The word error rate in this case was 67%. Note that this is better performance than in experiment 1. While it is difficult to gauge the statistical significance of this difference, nevertheless there are important differences between this experiment and experiment 1 that are worth noting: the speaker speaks very clearly, which is good for the acoustic models, and somewhat more sentences are used in this experiment to adapt the acoustic models. There are also no disfluencies in his speech, and he speaks in complete sentences, both of which are good for (i.e. correspond more closely to) the language model.

b) Unsupervised adaptation of the test utterances. In this experiment the first utterance was recognized, then the acoustic models were adapted based upon what was recognized. The resulting models were used to repeat the process on the next utterance until all 170 utterances were recognized. The word error rate dropped to 60%, a 10% improvement compared to a).

c) Supervised adaptation of the test utterances. In this experiment the first utterance was recognized, then the acoustic models were adapted based upon the correct, rather than the recognized, transcription. The resulting models were used to repeat the process on the next utterance until all 170 utterances were recognized. The word error rate dropped to 54%, a 19% improvement compared to a).

Table 2.

Experiment	Recognizer Accuracy (Word error rate)
No Additional Adaptation	67%
Unsupervised Adaptation	60%
Supervised Adaptation	54%

These results (summarized in table 2), while very preliminary in nature, are encouraging. Supervised adaptation gave the best result out of all of the experiments, while, even with poor recognition performance, unsupervised adaptation worked well enough to realize half the improvement made by supervised adaptation. Further research on a speaker adaptive system seems warranted based on these results, but also because of the usefulness of a speaker adaptive system: it could be applied to broadcast data with the same speakers, e.g. news anchors, hosts of talk shows, etc.

In many ways all three of these experiments point to the desirability of domain specific acoustic and language models. The more accurately the acoustic models were adapted, the more the recognition performance improved. There was less quantitative evidence regarding the language model, since no experiments were run exploring the benefits of adapting the Wall Street Journal language model to this domain. However, the perplexity of the narrator's test sentences was 532 using the Wall Street Journal language model, compared to a value of 235 using a test set of Wall Street Journal sentences. These numbers give a sense of how well the language model fits the test sentences, with a lower number indicating a better fit. Also, anecdotal examples of "Wall Street Journal like" phrases, such as "pork futures", being inserted in inappropriate places abound (see figure 1 for examples). Result c) suggests that with more data much better recognition accuracy is very likely (for the language models such data could come from the large quantity of transcripts of broadcast data which is available now). An important question to be addressed in future research is "how much data is needed to build good recognition models for this domain?"

Topic Identification Experiment

Given that the amount and variety of broadcast data constitutes a daily tsunami of information, it would be extremely useful to be

able to automatically sift through large amounts of data for subject, indexing it or for classifying it as "interesting" or "not interesting"

Figure 1.

Example 1. Wall Street Journal language model intrusion

Spoken: PROFESSORS STANLEY PONS AND MARTIN FLEISCHMANN
Recognized: PROFESSOR STARTING HOMES IN MODERATE INFLATION

Example 2. Typical recognition

Spoken: BUT OTHER SCIENTISTS BELIEVE THEY'VE CRACKED THE REPRODUCIBILITY PROBLEM
Recognized: THE OTHER SCIENTISTS BELIEVE THE FACTORY PRODUCES A PROBLEM

Example 3. Good recognition

Spoken: FOR THE CRITICS IT IS THE SHEAR DIVERSITY OF CLAIMS FOR EXCESS HEAT IN HEAVY WATER
Recognized: FROM CRITICS IT IS THE SHEAR DIVERSITY OF CLAIMS FOR EXCESS HEAT IN HEAVY WATER

based upon predetermined criteria. This could be a useful pre-processing step in the transcription process. Dragon developed a system for topic identification ([topic]) which was tested on the SWITCHBOARD corpus, and which has worked very well even when recognition performance, as measured by raw word correct, was very poor, a mere 22%. In addition to sorting data prior to transcription, such a topic identification system could be used to load an appropriate domain specific dictionary for the human user or domain specific language models for the recognizer during the transcription process. Given the current low level of recognition performance, the purpose of this experiment was to explore how well a topic identification system might work on broadcast data.

For this preliminary BBC topic identification test, the SWITCHBOARD-based system ([topic]) was used without modification. At the heart of this topic identification system are, for each pre-specified SWITCHBOARD topic, keyword lists and keyword probabilities. The keyword list for a given topic was, using standard statistical methods, selected from a training set containing SWITCHBOARD conversations on and off the topic. This training set of conversations was also used to estimate the probability the keyword occurs in

conversations on and off the topic. The system uses these keywords and probabilities to score a given conversation, on an unknown topic, against all of the pre-specified topics. These scores are then used to determine topic classification.

The recognition part of the experiment went as follows. The test set consisted of 200 utterances, divided into 20 units consisting of 10 utterances each. The units are the samples of television data which this experiment attempted to classify. Five of the units were from the news program and dealt with (the SWITCHBOARD topic) Crime. Two of the units, taken from the science program, described a cold fusion powered car and loosely correspond to the SWITCHBOARD topic Buying a Car. Thirteen units dealt with neither of these topics. After adapting the Wall Street Journal acoustic models using the 544 utterances that were not in the test set, the utterances in the test set were recognized. The recognition word error rate was 73%, similar to what was obtained in experiment 1.

The 20 units, assembled from the recognized speech, were then passed through the topic identification system described above. The best results were obtained on the topic Crime: 100% probability of detection at 13% false alarm rate. The Buying a Car topic did not fare as well: the two Buying a Car units ranked ninth and twelfth in the twenty units scored. There were too few units on the topic Buying a Car to attach too much significance to this result. Also, recall that the keywords used for scoring the messages were selected and trained using very different data, hence were not necessarily suitable for this test. For example, "japanese", "foreign" and "mine" are examples of keywords that were appropriate for the SWITCHBOARD task but that, in this experiment, resulted in strong scores for several off-topic units. In particular, "japanese" and "foreign" occurred frequently in the units taken from the science program in the contexts of investments. The news program units contained the keyword "mine" in the context "landmine". The Crime topic did not suffer as much from this defect since its keywords, such as "killed", are more universal signifiers. With keyword lists and probabilities trained from more appropriate data (large quantities of transcripts from other television programs, which are available now, would be suitable data), overall performance closer to the Crime topic could be expected for most topics in future experiments.

5. Future Directions for Research

Since large quantities of data will be needed to build good recognition

models for further explorations in broadcast data, a significant research problem is the automation of the data collection process. Broadcast audio data comes naturally in a long stream of many sentences (with some overlap between the speakers in some sorts of programming). To use this data for speech recognition experiments, the audio data needs to be processed into smaller sentence-like chunks and matching transcripts need to be generated. Simple speech chunkers are currently under development at Dragon which could assist in the data collection system. If verbatim transcripts are available, it might be possible to enhance the system to simultaneously chunk the transcriptions. It may, however, turn out to be simpler and cheaper to contract out the work of transcribing the chunked speech data.

After large amounts of broadcast data are collected, the first priority of any future work should be to get the recognition accuracy up to improved levels. As has been noted repeatedly in this report, it would be very interesting to build acoustic and language models directly from broadcast data and test their performance. The acoustic model building process can be simplified somewhat, since, based on the results of the segmentation experiment, the segmented acoustic data necessary to initiate the training process can be generated automatically using TIMIT models.

A serious issue that was not investigated in any of the experiments so far is how to deal with background noise, particularly music, in broadcast data. Nor were any experiments run to find out the extent to which background noise contributes to degradation in recognition performance. A suite of experiments to investigate these issues could be run along the following lines. Start with a set of noise free data, then create new sets of data by adding in various types noise at varying levels using standard sound mixing technology. Then acoustic models could be trained from each of these sets, or models could be trained from the data after noise cancellation or noise masking techniques were applied, and the results of recognition experiments could be compared. Since background noise is naturally present in most types of broadcast data, the lessons learned from these sorts of experiments would ultimately result in better recognition performance.

In the current transcription process, the transcriber often controls the playback of the audio tape with foot pedals while simultaneously typing. As was noted earlier, speech recognition technology could

simplify the transcription process by automatically synchronizing the playback of the audio with the typist's position in the transcript. A simple segmentation experiment was performed during the course of this investigation that worked out very well, suggesting that synchronization is plausible. But this segmentation experiment aligned the entire transcript to the acoustic data. A more interesting line of investigation for future work would be to develop algorithms to enable a recognizer to align a partial transcription to the acoustic data.

One possible scenario for an adaptive semi-automatic transcription system, mentioned above, is that the user corrects the output of a large vocabulary recognizer, with the recognizer adapting its models based upon the user's corrections. A more automated process would simply adapt the models based on what was recognized. A speaker adaptive version of these systems could be applied to a large collection of acoustic data generated by a single speaker. This sort of data is naturally generated by many television programs, e.g. a news program with the same announcer every day. A set of experiments along the lines of experiment 2 could be run, but with domain specific models and more data for recognition and adaptation, in order to get a better sense of how such systems might work. It would also be interesting to explore how well unsupervised adaptation works at various recognition error rates as well as, for supervised and unsupervised adaptation, how quickly and to what value the error rates converge.

Another proposed component of a semi-automatic transcription system is a topic identification system. This could be used to preprocess a huge amount of broadcast data by selecting only those segments on a given topic. During transcription it could be used to load topic specific on-line dictionaries for the human operator or topic specific language models for the speech recognizer. Since topic identification works well even when recognition accuracy is poor, both of these applications could be based upon a rapid but errorful recognition. The preliminary experiment in topic identification described above was successful enough to warrant further research. Given larger amounts of data, domain specific keyword lists and probabilities could be constructed to conduct experiments along the lines of those reported here. In addition, it would be very interesting to explore how much data need be sampled before a topic identification can be made reliably. It would also be interesting to run a suite of experiments to gauge the accuracy of the topic

identification system at several different recognition error rates.

6. Summary

The current transcription process is labor intensive and requires highly skilled, hence expensive, workers. The promise of state-of-the-art large vocabulary speech recognition technology is that, even though at present it cannot automatically transcribe speech very accurately, it could be part of a semi-automatic transcription system designed to allow less highly skilled workers to achieve the productivity of today's transcribers.

The BBC generously provided broadcast data which was used in several very preliminary recognition experiments. These experiments tested the data's usability and explored some possible uses for a speech recognizer in the transcription process. The recognition accuracy reported in these experiments was relatively low, but this is attributable to the fact that there was insufficient BBC data to fully adapt the acoustic models, built from the Wall Street Journal corpus, to the BBC domain. Promising strategies for improving performance were discussed, such as building models from broadcast data and experimenting with noise compensation. In the last experiment, a topic identification system was able to successfully classify the output of a speech recognizer against one topic, in spite of the errors in the recognized transcripts.

Based on what was learned from the interaction with transcription centers and the results of the initial experiments, future strategies for integrating large vocabulary speech technology in the transcription process were outlined. To give a few examples, a speech recognizer could be used to aid in the control of audio playback by aligning the audio data to what has been typed. Also, a typist and recognizer could work together to produce a transcript, with the user correcting the errors in the output of a large vocabulary continuous speech recognizer, and the recognizer adapting its models based on the user's corrections. Lastly, a topic identification system could automatically load an appropriate on-line dictionary. Much more research will be necessary to establish the feasibility of these strategies, but each strategy seems to be promising in light of the initial experiments.

References

[recog1] R. Roth, J. K. Baker, J. M. Baker, L. Gillick, M. Hunt, Y. Ito, S. Lowe, J. Orloff, B. Peskin, and F. Scattone, "Large Vocabulary Continuous Speech Recognition of Wall Street Journal Data", *Proc. ICASSP-93*, April 1993, Minneapolis, Minnesota, pp 640-643.

[recog2] L. Gillick, J. Orloff, R. Roth, F. Scattone and J. M. Baker, "Adaptation of Acoustic Models in Large Vocabulary Speaker Independent Continuous Speech Recognition", to appear in *Proceedings of the ARPA Spoken Language Systems Program*, March 1994.

[recog3] L. Gillick, J. Orloff, R. Roth, F. Scattone and J. M. Baker, "Studies in Large Vocabulary Speaker Independent Continuous Speech Recognition", to appear in *Proceedings of the ARPA Spoken Language Systems Program*, March 1994.

[IMELDA] M. Hunt, D. Bateman, S. Richardson, and A. Piau, "An Investigation of PLP and IMELDA Acoustic Representations and of their Potential for Combination", *Proc. ICASSP-91*, May 1991, Toronto, Canada, pp 881-884.

[topic] B. Peskin, L. Gillick, Y. Ito, S. Lowe, R. Roth, F. Scattone, J. M. Baker, J. K. Baker, J. Bridle, M. Hunt, J. Orloff, "Topic and Speaker Identification via Large Vocabulary Continuous Speech Recognition", *Proceedings of the ARPA Human Language Technology Workshop*, Plainsboro, NJ, March 1993, Morgan Kaufmann, pp 119-124.

[WGBH] *Real-Time Captioning and Real-Time Writing for Court Reporters*, The Caption Center, WGBH Educational Foundation, Boston, 1992.